Analysis of information literacy skills with data mining techniques

Phanisara Chanyam[1], Praweenya Suwannatthachote[2], Sungworn Ngudgratoke[3]

Abstract

The objective of this research was to classify the patterns of information literacy skills of Grade 12 students using data mining techniques. The research sample consisted of 384 Grade 12 students in Thailand in the academic year 2021, who had a confidence level of 95%. The data analysis was divided into two steps: The first step was clustering to group the information literacy skills of the students with similar characteristics using the K-means method, which resulted in three clusters based on three indicators: 1) Information sources and information resources, 2) Information retrieval strategies, and 3) Ethical and legal use of the information. The second step was classification to predict the patterns of the information literacy skills of the Grade 12 students obtained from the clustering, which had a total of 74 conditions.

Keywords: Information literacy skills, Data mining, Cluster, Classification, K-means, Decision tree

(Received: 26 April 2023; Revised: 11 December 2023; Accepted: 11 December 2023)

[1] Ph.D. student in Educational Technology and Communications, Faculty of Education, Chulalongkorn University email: phanisara13@gmail.com

[2] Associate Professor, Department of Educational Technology and Communications, Faculty of Education, Chulalongkorn University email: praweenya.s@chula.ac.th

[3] Associate Professor, Educational Measurement and Evaluation Program, School of Educational Studies, Sukhothai Thammathirat Open University email: sungworn.ngu@stou.ac.th

Introduction

The flow of data from various sources in the present time is massive and fast-paced, thus allowing individuals to access information constantly in various formats, including text, images, audio, and videos. These data may come in diverse mixed media forms through various channels. All the information enters people's perception from the moment they wake up until they go to sleep through multiple avenues, such as street media, radio, television, and internet-based media like websites, blogs, and social media. However, the majority of accessible information often lacks filtration raising concerns about its accuracy (Maitaouthong, 2018). Therefore, it would be crucial to prepare young people with the resilience to receive and utilize information resources, especially electronic resources. Hence, students would need to possess information literacy skills as a foundation to ensure efficient use of electronic information resources (Odede & Nsibirwa, 2018). Moreover, information literacy skills would be crucial for learning in the 21st century (Maitaouthong, 2018) to foster lifelong learning, especially for students transitioning into higher education levels where they would need to apply information literacy skills in their studies and daily lives, which would require more specific and specialized conditions. This would enable learners to use information appropriately because information literacy would be the ability to identify, locate, evaluate, and effectively use information when needed. As such, this would serve as a fundamental skill across all disciplines, learning environments, and educational levels, as well as be a lifelong learning skill that could empower learners to take ownership and control over their learning (Arya, 2014).

Information literacy skills do not come naturally to humans but are developed through learning and practice (Bhornchareon et al., 2020). Additionally, information literacy enhances learners' self-directed learning opportunities and expands learning beyond the traditional classroom setting because it empowers learners with the ability to assess, manage, and utilize information beyond classroom learning. This would be crucial because learners are surrounded by an immense amount of information (Association of College & Research Libraries [ACRL], 2000). However, Sacchanand (2018) found that the Thai education curriculum separated information literacy skills into fragmented components rather than integrated them into the information process. In Thailand, information literacy skills are partially included in the core curriculum of basic education, specifically within the subjects of technology (computing) and science, which have learning standards and indicators related to information literacy skills, media, and digital literacy. This is also consistent with Saechan and Siriwipat's study (2017), which found that the information literacy skills of Islamic private high school students in the southern region of Thailand were categorized into five levels: not passing, passing, average, high, and the highest. The information skills that students did not pass, included the ability to analyze, evaluate, and select the desired information, as well as the knowledge and skills necessary for using information technology in various forms. This aligned with Pawinun's findings (2022) that secondary schools in the Bangkok metropolitan area and provincial schools had significant differences in information literacy skills and technology. This had a significance level of 0.05. This indicated the variations in readiness for teaching and learning management. Nevertheless, further research found that high school students had high-level information literacy skills with significant differences between lower secondary and upper secondary students at a significance level of 0.05 due to the differences in the content scope of the subjects. In addition, the evaluation of information literacy skills could vary in the format and data collection, thus making it difficult to compare the information literacy levels of students. Moreover, when considering the information literacy levels of graduate students from various universities, according to the research of Bhornchareon et al. (2020) and Hussakhun and Sirichote (2020), students had moderate-level information literacy skills. Bhornchareon et al. (2020) also found that students had low-level information literacy skills in terms of accessing and evaluating information.

Due to the significance of information literacy skills among upper secondary school students as a foundation for further education at the higher education level, establishing a strong foundation or promoting students' knowledge, understanding, and higher-level information literacy skills would be crucial for their future learning and life. Students come from diverse school types, and the specific content scope of the completed curriculum varies. Therefore, this research classified the information literacy skills of Grade 12 students in Thailand in order to design a learning system that would enhance the necessary information literacy skills for their future education at the most suitable higher education level.

Literature review

1. Information Literary

This research utilized the "Framework for Information Literacy for Higher Education" developed by the Association of College and Research Libraries (2016) as a framework for the study.

"Information literacy is the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning."

1) Authority is constructed and contextual

Information resources reflect the expertise and credibility of their creators, and are evaluated based on the desired information and the context in which they would be used. The authority generated in different communities could lead to different types of power, which could help determine the level of authority required for a particular information context.

2) Information creation as a process

Information in various formats would be created to be transmitted through selected methods. The iterative process of inquiry, creation, evaluation, and dissemination of information would take various forms, and the outcomes would reflect the diversity of the information.

3) Information has value

Information possesses multiple dimensions of value, including commercial, educational, motivational, negotiation, and societal understanding. Legal and socio-economic factors would also influence the creation and dissemination of information.

4) Research as inquiry

Research involves iterative processes and is dependent on increasingly complex questions that would lead to further inquiry or investigation in different directions.

5) Scholarship as conversation

The academic and research community actively contributes to discourse by supporting and disseminating new findings and developments over time from diverse perspectives and interpretations.

6) Searching as strategic exploration

Information seeking is not a linear process and often requires multiple iterations. This would require broad evaluation from multiple sources and flexibility in navigating the main pathways while developing new understandings.

2. Data Mining

Data mining is a term that illustrates the process of extracting knowledge from vast amounts of data. The result of data mining is knowledge, which is obtained through analyzing data to uncover hidden patterns (models) that depict the characteristics and relationships within the data. These patterns reflect recurring occurrences (repeat), thus enabling prediction. This acquired knowledge could be utilized in various fields, such as medicine, business, and education (Auwatanamongkol, 2020; Limpiyakorn, 2008).

Types of data mining methods:

1) Predictive data mining: This method is used to predict future outcomes based on inferences drawn from the current data set. It would employ supervised learning, where the model would be the outcome of data mining. The data would be categorized into known groups based on class labels and utilize training data; for example, in classification tasks (Auwatanamongkol, 2020; Limpiyakorn, 2008).

2) Descriptive data mining: This method aims to discover patterns within a given data set. It would not require training data and would employ unsupervised learning. Descriptive data mining involves identifying associations and clustering data to uncover hidden relationships, factors, or causes (Auwatanamongkol, 2020; Limpiyakorn, 2008).

3. Data classification

1) Classification aims to create a model for predicting class labels. It would use categorical data and train the classifier to learn classification from examples in the training data set. The model would be evaluated for its performance and generalization using the test data set. Once the model is obtained, it could be used to classify new data whose class labels are unknown (Auwatanamongkol, 2020; Limpiyakorn, 2008).

2) Clustering involves grouping data where the most similar data points would be placed in the same cluster, while different clusters would be distinct. The number and types of clusters would be unknown in advance. Clustering is used to understand data distribution or serve as a preprocessing step for other algorithms (Limpiyakorn, 2008).

Research Aims

The aim of this research was to classify the patterns of information literacy skills of Grade 12 students.

Research Methodology

Research Sample and Sampling

The population consisted of Grade 12 students in Thailand.

The sample included Grade 12 students selected specifically based on the criteria set by each region's schools and the analysis of the class teacher. The sample size was determined to be 384 students, which represented a 95% confidence level from the total population of 233,632 students in Thailand, academic year 2021. Data was collected from 18 schools nationwide, consequently resulting in a sample size of 396.

Research Tool

The research tool used in this study was a self-developed information literacy skills assessment questionnaire. It was divided into two parts: 1) general information and 2) questions on information literacy skills. The questionnaire was based on the analysis of the "Framework for Information Literacy for Higher Education" by the Association of College and Research Libraries (2016) comprising a total of six measurement indicators:

1. Information sources and information resources
2. Information retrieval strategies
3. Information evaluation
4. Information analysis and interpretation
5. Appropriate referencing of information
6. Ethical and legal use of information

The questionnaire underwent quality checks, including item content validity and language accuracy, measured through the calculation of Item Objective Congruence (IOC), where questions with IOC values between 0.60 and 1.00 were considered suitable. The revised questionnaire was then piloted on a sample of 150 participants, who were similar to the group under study. The selection of exam questions was based on the analysis of item difficulty, which ranged from 0.00 to 1.00, with lower values indicating more difficult questions. In this study, a difficulty criterion between 0.21 and 0.80 was used. Exam questions with higher values indicated easier questions. Additionally, the discrimination power of the questions was analyzed with discrimination values ranging from -1.00 to 1.00. Questions with negative values were discarded since they had the opposite effect, thus meaning that less competent individuals could answer correctly while more competent individuals could not. Questions with discrimination values of 0.20 or higher, hence indicating good group discrimination, were selected. Furthermore, questions with difficulty values outside the established range and those with poor discrimination power were excluded. After analyzing the test data, the standard deviation was found to be 4.91, and the reliability coefficient (KR-20) was 0.7919. Therefore, it was recommended to retain 25 out of the total 36 exam questions as they exhibited good qualities.

Research Process

The research process consisted of the following steps:

Step 1: Clustering to group the information literacy skills of Grade 12 students with similar characteristics.

Step 2: Classification to predict the patterns of information literacy skills of Grade 12 students obtained from the clustering.

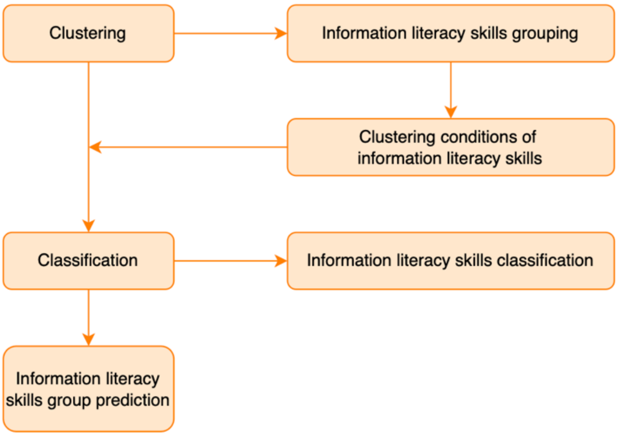

Clustering
Information literacy skills grouping
Clustering conditions of information literacy skills
Classification
Information literacy skills classification
Information literacy skills group prediction

Figure 1: Depicts the data mining analysis process.

Data Analysis

Step 1: Cluster analysis was conducted to analyze the grouping of the data with similar characteristics but distinct from other groups, in order to study the specific characteristics of each group. The K-means method was utilized for this research with a sample size of over 200 samples.

In this step, RapidMiner Studio (version 9.10.001) was employed for the data analysis by utilizing the following process:

1) Data preparation phase: Data from the information literacy skills questionnaire (data set) were collected to assess the quality. The data were selected, checked for completeness and accuracy, and then refined.

1.1) Data cleaning involved removing missing data, such as empty values or incomplete data.

1.2) Data transformation involved converting the data to be compatible with the model, such as transforming text data into numerical data to be suitable for analysis using the algorithms. The data were transformed into a CSV format, which was a standardized file format for tabular data. The scores of the six information literacy skills indicators were transformed using z-transformation because the maximum scores for each indicator were not equal. However, upon analyzing the data from the six indicators, it was found that the data had a high level of dispersion, consequently making it difficult to identify clear groups. Therefore, the researchers combined the scores of the six indicators into three dimensions based on the related content (Table 1).

Table 1: Combination of the scores from the six indicators into three dimensions.

Indicators	Code
1. Information sources and resources	IL1
2. Information searching strategies	IL2,3,4
3. Information evaluation	
4. Information analysis and interpretation	
5. Information citation	IL5,6
6. Information ethics and law	

2) Modeling phase: Using the K-means technique with RapidMiner Studio, the data were analyzed to group the information literacy skills obtained from the clustering (Figure 2).

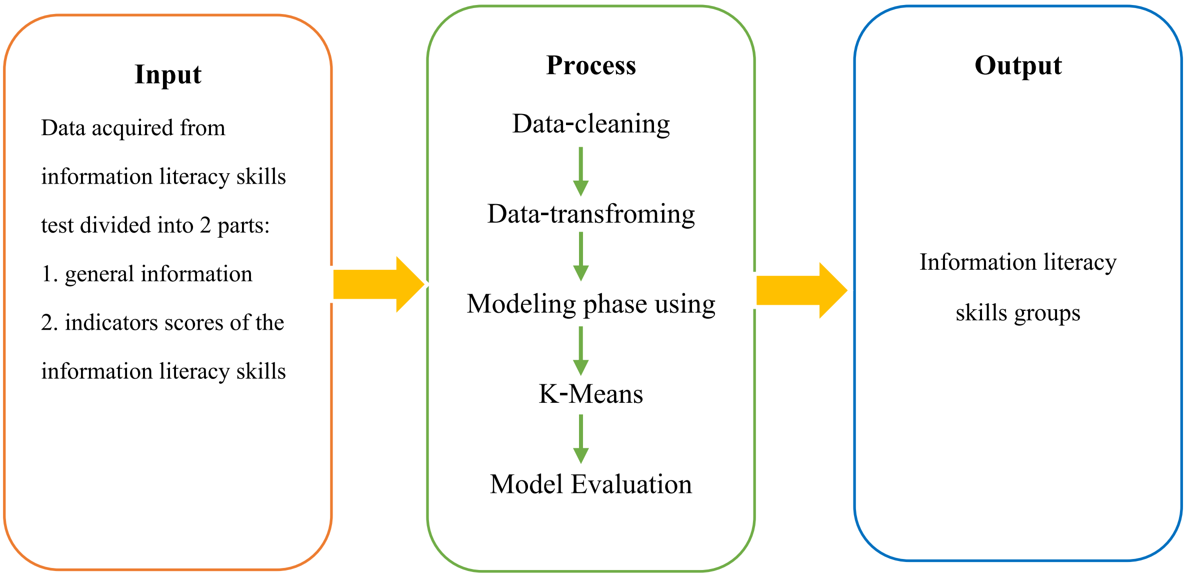


Figure 2: Cluster analysis process using K-means

Step 2: (Classification) To predict the patterns of information literacy skills among Grade 12 students, the data obtained from the analysis and the models derived from the clustering process were used. The analysis was performed using RapidMiner Studio (version 9.10.001) software by utilizing the following process:

1) Data preparation phase: The scores from the information literacy skills questionnaire (data set) were used to assess the quality of the data. Data selection, completeness checking, and data accuracy were conducted before further analysis. The obtained data were then refined.

1.1) Data cleaning: Missing data were removed.

1.2) Data preparation: In this step, the data obtained from the clustering process in Step 1 were used as a condition for evaluating the patterns. The information from the three dimensions of the information literacy skills was considered using the scores as a basis for evaluation. The data were transformed into a table format with intervals (Table 2).

Table 2: Conditions for evaluating the transformed score data from the clustering process in Step 1.1

IL1 (I: Information)		**IL2-4 (S: Strategy)**		**IL5-6 (E: Ethics)**	
Score (5)	**Code**	**Score (10)**	**Code**	**Score (10)**	**Code**
0-1	IF	0-1	SF	0-1	EF
2	ID	2-3	SD	2-3	ED
3	IC	4-5	SC	4-5	EC
4	IB	6-7	SB	6-7	EB
5	IA	8-10	SA	8-10	EA

2) Modeling phase: In this step, a decision tree modeling technique was used to make decisions. The process was as follows:

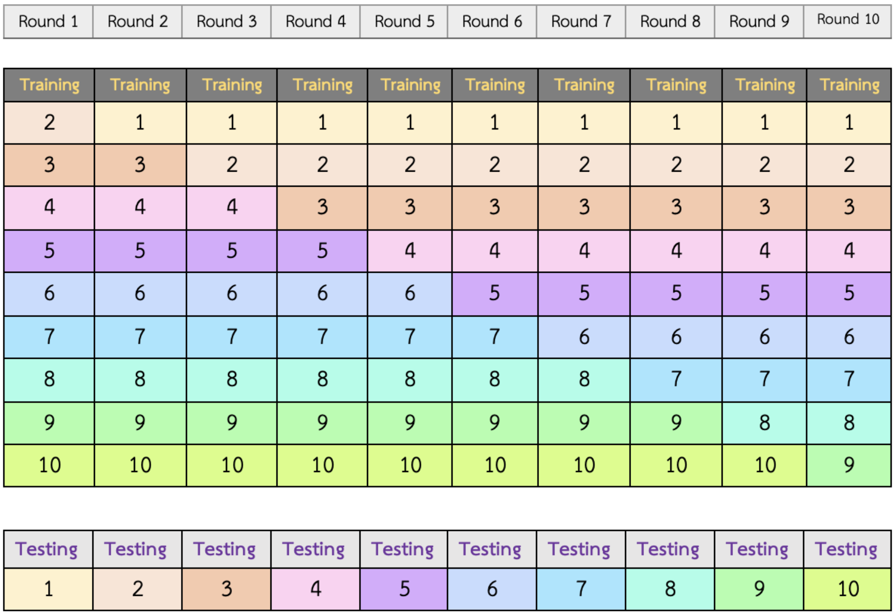
2.1) Representation of the data based on the characteristics of the information literacy skills clusters was obtained from clustering (Table 3).

Table 3: Example of data representation for the information literacy skills clusters

No.	Cluster	IL1	IL2-4	IL5-6
1	Low	IF	SF	EB
2	Low	IF	SD	EC
3	Medium	IF	SC	EB
4	High	IB	SB	EB
5	Medium	IC	SD	EB

2.2) Creation and testing of the prediction models was conducted by dividing the data into a training data set and a test data set. The goal was to find groups or patterns (models) of information literacy skills. K-fold cross validation was used, specifically 10-fold cross validation, which involved the following process:

The data were divided into 10 equal parts. After that, the model's performance was tested 10 times (Figure 3).



Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10
Training	Training	Training	Training	Training	Training	Training	Training	Training	Training
2	1	1	1	1	1	1	1	1	1
3	3	2	2	2	2	2	2	2	2
4	4	4	3	3	3	3	3	3	3
5	5	5	5	4	4	4	4	4	4
6	6	6	6	6	5	5	5	5	5
7	7	7	7	7	7	6	6	6	6
8	8	8	8	8	8	8	7	7	7
9	9	9	9	9	9	9	9	8	8
10	10	10	10	10	10	10	10	10	9
Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing
1	2	3	4	5	6	7	8	9	10

Figure 3: 10-Fold cross validation

2.3) The performance of the models was evaluated using 396 sets of the test data. The accuracy of the classification models was measured to assess their effectiveness.

Findings

The use of data mining techniques to predict the information literacy skills of Grade 12 students provided the following results:

1. The clustering results were conducted to group the information literacy skills of 396 Grade 12 students based on the test scores. The test results revealed that using all six indicators of the information literacy skills from this research did not yield clear and distinct groups due to excessive data granularity. Therefore, the researchers combined the closely related indicators into three dimensions: 1) information sources and information resources, 2) information retrieval strategies, and 3) ethical and legal use of the information. These dimensions allowed for the grouping of the information literacy skills into three clusters, which were the most suitable and appropriate for dividing the skills (Table 4).

Table 4: The number of Grade 12 students in each cluster and their respective patterns of information literacy skills.

Cluster	Level	N	%
0	High	77	19
1	Low	172	43
2	Medium	147	37
Total		396	100

From the grouping of the information literacy skills in each dimension, the details are shown in Figure 4.

From Figures 1, 2 and 3, it could be observed that through the pairing and comparison of each dimension of information literacy skills, the following patterns were found:

1) Cluster 1 exhibited the lowest scores in the dimension of the information sources and information resources although some parts fell within the moderate to high range. When clustered, this remained categorized as a group with information literacy skills at a low level. On the other hand, Cluster 0 demonstrated moderate to high scores in both the information sources and information resources (il1) and information retrieval strategies (il2, 3, 4), thereby indicating a group with high-level skills. Cluster 2 belonged to the moderate-level skills group, as the scores fell within the middle range between the low and high clusters.

2) Cluster 1 exhibited the lowest scores in the dimension of the information sources and information resources although some parts fell within the high range. When paired with the ethical and legal use of the information (il5, 6), it was found that the scores were at a low level. On the other hand, Cluster 0 demonstrated moderate to high scores in both the information sources and information resources (il1) and ethical and legal use of information, thus falling within the low to high range. This group could be considered to have overall high-level skills. Cluster 2 belonged to the moderate-level skills group, as the scores fell within the middle range between the low and high clusters.

3) Cluster 1 had low to moderate scores in information retrieval strategies (il2, 3, 4), but when paired with the ethical and legal use of the information (il5, 6), it was found that the scores were at a low level. On the other hand, Cluster 0 had moderate scores in both the information retrieval strategies and ethical and legal use of the information, hence falling within the high range. This group could be considered

to have overall high-level skills. Cluster 2 fell into the group with moderate-level skills, as the scores were in the middle range between the low and high clusters.

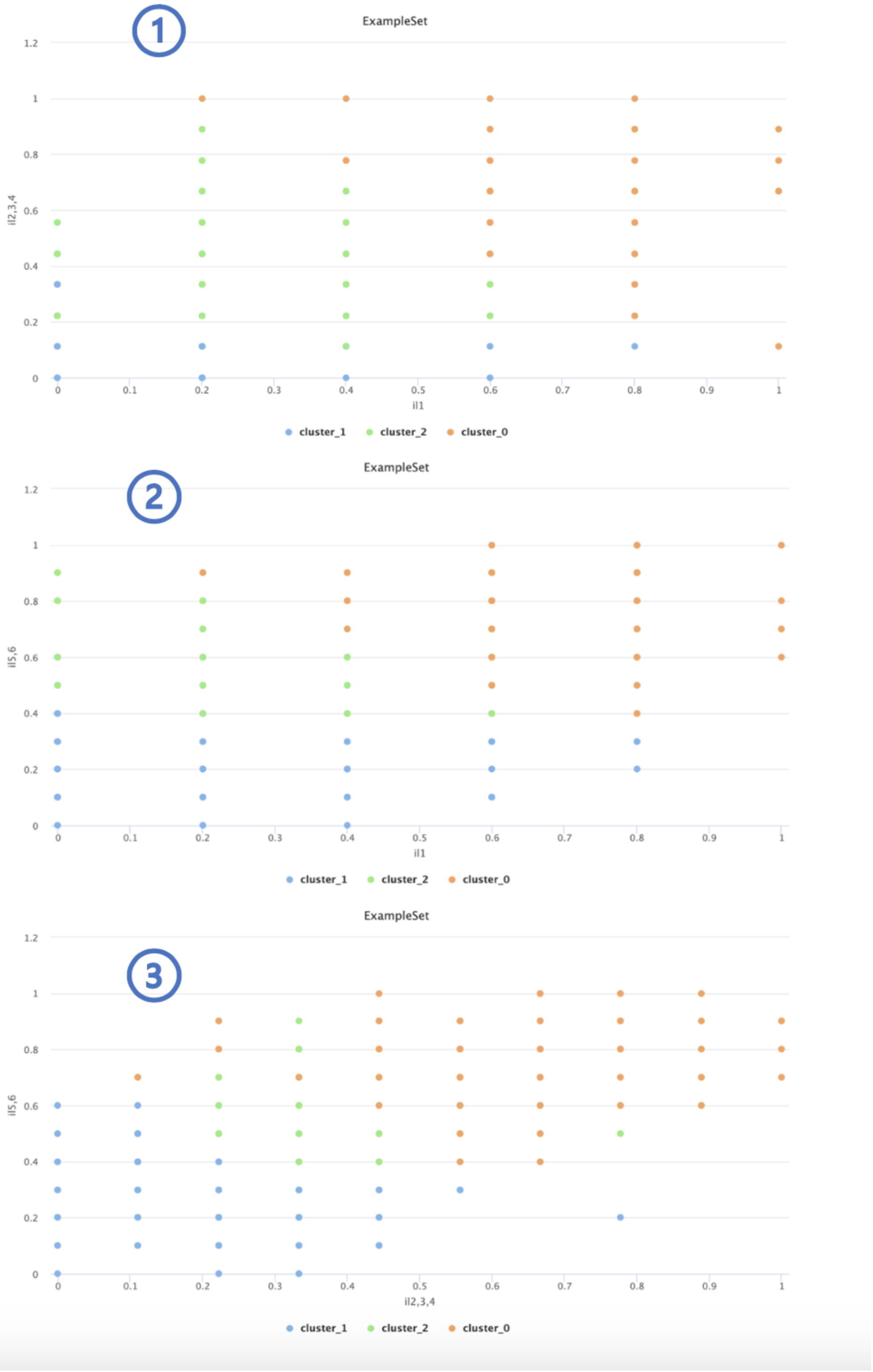
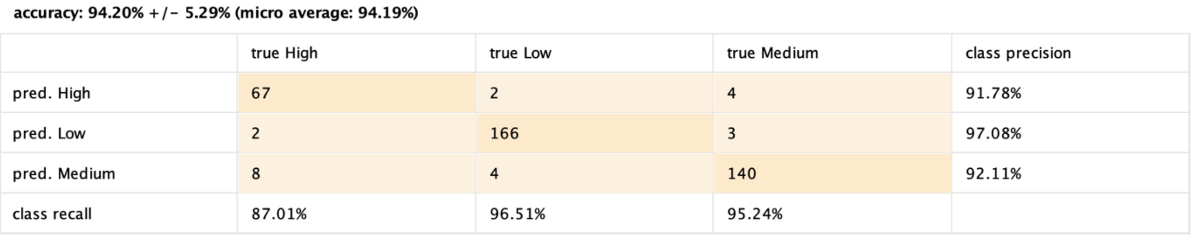


Figure 4: Clusters of the information literacy skills of Grade 12 students

In summary, from the clustering analysis, it could be observed that Cluster 1 had low scores in all three dimensions on average even though some dimensions may have higher scores. Thus, individuals scoring in Cluster 1 could be considered to have low-level information literacy skills. Conversely, Cluster 0 had relatively high scores in all dimensions. Although there may be low scores in some dimensions, no dimension scored zero. Therefore, individuals scoring in this cluster could be considered to have high-level information literacy skills. Cluster 2 represented a group with moderate-level information literacy skills, as this combined low, moderate, and high scores in different dimensions without extreme lows or highs simultaneously.

2. The results of the classification for predicting the patterns of the information literacy skills of Grade 12 students achieved an accuracy rate of 94.19% for the model (Figure 5).



accuracy: 94.20% +/- 5.29% (micro average: 94.19%)

	true High	true Low	true Medium	class precision
pred. High	67	2	4	91.78%
pred. Low	2	166	3	97.08%
pred. Medium	8	4	140	92.11%
class recall	87.01%	96.51%	95.24%	

Figure 5: Accuracy of the model in the classification process.

The analysis of the patterns of the information literacy skills of the students using the decision tree technique revealed a total of 74 conditions. These conditions were divided into three groups: weak (29 conditions), moderate (20 conditions), and strong (25 conditions) (Figure 6).

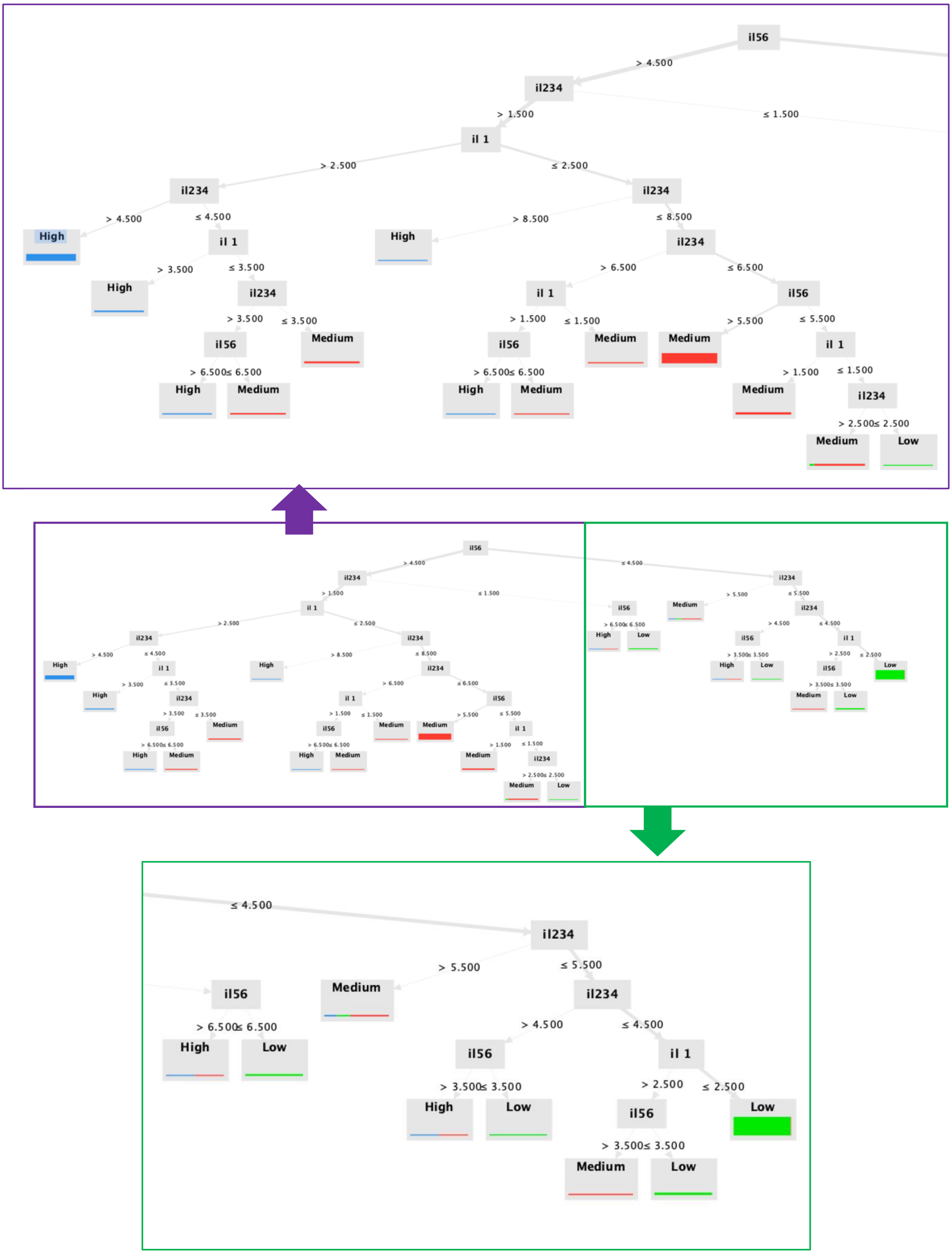


Figure 6: Decision tree

Table 5: Example of the patterns of the information literacy skills of Grade 12 students.

IL 1	IL 2,3,4	IL 5,6	Class
IF	SB	EC	Med
ID	SF	EF	Low
		ED	Low
		EC	Low
IC	SC	EF	Low
		ED	Low
		EC	Med
		EB	Med
		EA	High
IB	SF	EF	Low
	SB	EC	High
IA	SF	EB	High
		EA	High

From the example in Table 5, it could be observed that there were many conditions for classifying the information literacy skills of Grade 12 students. For instance, in the case of the "Low" class label, the conditions for classification were il1 = IF, il2, 3, 4 = SB, and il5, 6 = EF. These conditions corresponded to the score ranges of il1 = 0-1, il2, 3, 4 = 6-7, and il 5, 6 = 4-5.

Conclusions and Recommendations

The research findings revealed that Thai Grade 12 students' information literacy skills varied across three indicators: information sources and information resources, information retrieval strategies, and ethical and legal use of information. Data mining analysis uncovered different perspectives and relationships within the data. Upon analyzing the clustering results, it was observed that students had varying levels of knowledge, understanding, and information literacy skills, thus resulting in three distinct groups. Further examination revealed that all three groups scored the highest in the ethical and legal use of information compared to the other two aspects. To utilize the grouping data, this research analyzed the conditions of each group obtained from clustering and used them as attributes for data prediction. The data mining technique employed was classification with the decision tree being the most suitable method considering the limitations and specifics of the research data. The analysis yielded a total of 74 prediction conditions. However, to ensure a comprehensive prediction of the students' information literacy skills in the future, it would be necessary to consider all possible events based on the conditions of il1, il2, 3, 4, and il5, 6 which would amount to five conditions each, as mentioned earlier. This totaled 53 events or 125 conditions.

This research demonstrated the potential application of data mining techniques to uncover hidden relationships or specific data within a data set, such as clearly defined test scores or information that teachers would be interested in or need to know. These findings could be utilized in instructional design or future evaluations. However, when using data mining for data analysis, it would be important to consider the characteristics and appropriateness of the data, as data may vary in nature and therefore require different methods. Furthermore, the findings of this research could be used as a guideline for developing information literacy skills among Grade 12 students who would be transitioning to higher education. This aimed to prepare them for undergraduate studies and could be valuable for university instructors or stakeholders involved in designing instructional strategies or curriculum enhancement activities to fully develop students' information literacy skills. This would serve as an initial step in understanding students' limitations or strengths, thereby aiding instructors in designing learning objectives, activities, or teaching methods that would best suit the students. Additionally, the framework for information literacy skills outlined by the Association of College and Research Libraries (2016) provided a foundation for the appropriate skills and abilities that students at the higher education level should possess. These skills would be crucial for achieving greater success in learning, particularly as university-level education demands more profound information literacy skills. As a consequence, developing these skills would be essential for academic pursuits, work environments, and daily life, as they would prepare individuals for future societal engagement.

References

Arya, S. (2014). Information literacy programmes and practices: A survey of selected higher institutions of Udaipur district. **Global Journal of Academic Librarianship**, 1(1), 9-18. https://www.ripublication.com/gjal/gjalv1n1_02.pdf

Association of College and Research Libraries. (2000). **Information literacy competency standards for higher education**. The Association of College and Research Libraries. http://hdl.handle.net/11213/7668

Association of College and Research Libraries. (2016). **Framework for information literacy for higher education**. The Association of College and Research Libraries. https://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/infolit/framework1.pdf

Auwatanamongkol, S. (2020). **Data Mining** (3rd ed.). National Institute of Development Administration Press. [In Thai]

Bhornchareon, S., Techataweewan, W., & Prapinpongsakorn, S. (2020). Factors influencing information literacy among undergraduate students of Rajamangala University of Technology Phra Nakhon. **T.L.A. Research Journal**, 13(2), 58-76. https://so06.tci-thaijo.org/index.php/tla_research/article/view/246906 [In Thai]

Han, J., Kamber, M., & Pei, J. (2012). **Data mining: concept and techniques** (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Hussakhun, D. & Sirichote, P. (2020). Information literacy of undergraduate students at Rajabhat Universities in the northeast region. **Journal of Graduate School Sakon Nakhon Rajabhat University**, 17(76), 238-244. https://so02.tci-thaijo.org/index.php/SNGSJ/article/view/211905 [In Thai]

Limpiyakorn, Y. (2008). **Data Mining**. [Unpublished manuscript]. Department of Computer Engineering, Chulalongkorn University. [In Thai]

Maitaouthong, T. (2018). Information literacy skills for the 21st century learning. **Journal of Humanities and Social Sciences, Phranakhon Si Ayutthaya Rajabhat University**, 6(2), 171-189. https://so03.tci-thaijo.org/index.php/husoarujournal/article/view/260329 [In Thai]

Odede, I. R. & Nsibirwa, Z. (2018). Information literacy skills in using electronic information resources. **Library Philosophy and Practice (e-journal)**. Article 1947. http://digitalcommons.unl.edu/libphilprac/1947

Pawinun, P. (2022). Guidelines for information and digital literacy skills development of secondary education students. **T.L.A. Bulletin**, 66(1), 87-103. https://so06.tci-thaijo.org/index.php/tla_bulletin/article/view/253225 [In Thai]

Sacchanand, C. (2018). Development of information literacy promotion model for high school students. **T.L.A. Research Journal**, 11(2), 45-60. https://so06.tci-thaijo.org/index.php/tla_research/article/view/150511 [In Thai]

Saechan, C. & Siriwipat, V. (2017). Information literacy of upper-secondary students in the Islamic private schools in southernmost provinces. **Library and Information Science Srinakharinwirot University**, 10(2), 32-48. https://ejournals.swu.ac.th/index.php/jlis/article/view/9955 [In Thai]